Investigating the relationship between physical activity measured by hip-worn accelerometers and mortality

by

Yeya Zheng

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

May, 2018

Abstract

The associations between physical activity, demographic, and health-related characteristics and 5-year mortality were studied using the National Health and Nutrition Examination Survey (NHANES). Forward selection revealed that age, smoking status, drinking status, gender, comorbidity information including whether the person has CHF, cancer, stroke or diabetes at current age, blood pressure difference, and one fragmentation metric that quantifies the frequency of transitioning from active to sedentary are the most predictive predictors for 5-year all-cause mortality. These risk factors were used to build a prediction model, which has a 10-fold cross-validated AUC equal to 0.828. Our competing risk analyses for heart disease and cancer related deaths provided further insights into cause-specific predictors. In general, the results are consistent with our findings when all cause mortality is of concern, however, there are discrepancies which might be due to the decrease in number of events from all cause to cause specific mortality data. We conducted an upstrap re-sampling analysis to evaluate the effect of sample size and event per variable(EPV) on the power to detect the significances of regression

coefficients. In general, the power to detect the significant effect for each covariate goes up as the sample size and EPV increases. Moreover, we built a Shiny APP to translate our work to the general public. In conclusion, our research demonstrates the beneficial health effects of physical activity, non-smoking, and moderate drinking.

Primary Reader: Ciprian Crainiceanu

Secondary Reader: Mei Cheng Wang

Acknowledgments

I would first like to thank my thesis advisor Dr.Crainiceanu Ciprian for his patience, motivation, and immense knowledge. His guidance helped me to conduct this research and write this thesis.

I would also like to acknowledge Dr. Wang Mei-Cheng who was the second reader of this thesis, and I am gratefully indebted to her for her very valuable comments on this thesis.

I would also like to thank Dr. Zipunnikov Vadim, Andrew Leroux, Junrui Di, and Jiawei Bai, and the other members of the Wearable and Implantable Technology (WIT) group, for their help and input with this research project.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The National Health and Nutrition Examination Survey (NHANES) is a cross-sectional, nationally representative survey designed to evaluate the health and nutritional status of adults and children in the United States [2]. The survey samples around 5000 non-institutionalized civilians annually to represent the US population. In particular, NHANES oversamples underrepresented groups, including elderly people 60+ years old, African American, Asian, and Hispanic. The survey involves a 4-stage process to sample participants, which indicates that the sample is not a simple random sample from the US population. To make the sample representative for the US population each individual sampled in the NHANES has a survey weight, which is defined as the number of individuals in the US population represented by that individual. These survey weights need to be incorporated in any analysis to ensure that results are generalizable to the US population. The

survey collects demographic, socioeconomic, dietary, and health-related information through home interviews, and medical, dental, physiological measurements through physical examinations in mobile centers[2]. Moreover, NHANES started to monitor participants' physical activity through accelerometer during a 1-week study for its 2003-2004 and 2005-2006 cohorts. The National Center for Health Statistics also provides a mechanism for linking NHANES waves with death certificate records from the National Death Index (NDI)[9]. This allows us to investigate the associations between participants' activity and other non-activity related characteristics and future mortality.

For our research, we are interested in: 1) exploring the associations between participants' physical activity, demographic, and health-related characteristics and 5-year mortality; 2) identifying the ranking of the most predictive predictors and their relative effects on mortality; and 3) and building prediction models to predict 5-year mortality. The prediction model was developed into a Shiny APP (`https://yezheng42.shinyapps.io/MortalityCalculator/`) that takes users' input about their demographic, physical activity, and health-related information, and then predict user's 5-year all-cause mortality and relative risk of death as compared with a reference individual . Users can change modifiable health risk behaviors and see how it impacts their predicted mortality over time. The APP has the potential to translate our research work to the general public, and it also provides translational individualized advice. In addition to the overall mortality,

we have also examined the associations between participants' physical activity, demographic, and health-related characteristics with heart disease-specific mortality and cancer-specific mortality through competing risk models.

The paper is structured as follows: In Chapter 2, we will briefly introduce the NHANES dataset and the accelerometer metrics we derived. In Chapter 3, we present some exploratory plots to investigate the associations between physical activity and mortality, correlations between activity metrics, and the potential modifying effect of age on the association between physical activity and 5-year all-cause mortality. In Chapter 4, we identify the ranking of predictors and select the best-performed prediction model for 5-year mortality. In Chapter 5.1 we discuss our competing risk analysis and in Chapter 5.2 we introduce a bootstrap method we used to assess the impact of sample size on the power of regression coefficients for competing risk model, and report the our simulation results. In Chapter 6 we introduce our Shiny APP, and finally in Chapter 7 we conclude with a brief discussion on the results.

Chapter 2

Data

Originally, there are 14,631 participants in NHANES 2003-2004 and 2005-2006 cohort. We excluded participants who: 1) were younger than 50 y.o. or older than 85 y.o. at the time they wore the accelerometer (9369 participants) 2) had fewer than 3 days of valid data, either due to less than 10 hours of estimated wear time or due to either of the two NHANES provided quality flags indicate data quality concerns (1832 participants) 3) were ineligible for mortality follow-up due to insufficient identifying data to create a National Death Index submission record (6 participants) 4) died from accidents (8 participants) 5) were alive, but had a follow up less than 5 years or died after 5 years (183 participants) 6) who did not have the cause of death available (1 participant) or 7) had missing any of the demographic predictor variables we adjust for (261 participants). Eventually, there were 2971 participants from the NHANES 2003-2004 and 2005-2006 cohorts for my analysis.

The accelerometer data are originally in long format, with one row per subject-minute. Using the data package(nhanesdata)[7] we transformed the accelerometer data into the standard 1440+ format. In this format each row corresponds to per subject-day and each cell corresponding to minute-level activity count. Figure 1 displays an example of the processed accelerometer data format. We have also created a data matrix of non-wear flags using established algorithms to account for the non-wear time [10]. The data matrix is also in the 1440+ format, and each cell contains a binary outcome with 1 flagged for non-wear and 0 flagged for wear. Wear/non-wear flags were estimated from the data using the standard NHANES algorithms.

Starting from the processed accelerometer data and the flag data matrix, we have derived the activity metrics using the procedure listed in Table 1. Suppose that i=1, 2, ..., N denotes number of subjects and j=1, 2, ..., M denotes the number of days for each subject, and t=1, 2, ..., 1440 denotes minutes on each subject-day. Let AC_{ijt} be the activity count for subject i on day j minute t, and Z_{ijt}be the non-wear flag for subject i on day j minute t. We use the following notation:

$$Z_{ijt} = \begin{cases} 1 & nonwear \\ 0 & wear \end{cases} \tag{2.1}$$

The cut-point for sedentary time is set to be 100 and the cut-point for moderate-to-vigorous(MVPA) activity is set to be 2020. That is, any minute with activity count above 2020 is accumulated into MVPA time and any minute with activity

Figure 2.1: Example of the processed accelerometer data format

	SEQN	MIN1	MIN2	MIN3	.....	MIN1440
1	21009	0	0	0	.....	0
2	21009	0	0	0	.....	0
	.....	.....	.....	.....	.....	.....
7	21009	0	0	0	.....	0
772	21868	0	0	0	.....	0
773	21868	0	0	0	.....	0
	.....	.....	.....	.....	.....	.....
778	21868	0	0	0	.....	0

count below 100 is accumulated into sedentary time.

For fragmentation metrics, we will introduce a different system of notation. Let D_A denote the duration of the longest active bout, $n_A(t)$ denote the number of bouts of length t, and $n_A^c(t)$ denote the number of active bouts of length $\leq t$. Then the total active time can be represented as $T_A = \sum_{t=1}^{D_A} n_A(t) * t$, and the total number of active bouts can be represented as $n_A = \sum_{t=1}^{D_A} n_A(t)$. Notations for sedentary bouts are similar except for all subscripts changed to "S".

Table 2.1: Activity metrics along with the procedure used to derive them and their interpretation

Metrics	Interpretation	Derivation
$\overline{TLAC_i}$	Average total log activity count for subject i	$\frac{1}{m}\sum_{j=1}^{m}\sum_{t=1}^{1440} log(AC_{ijt}+1)$
$\overline{Pdaytime_i}$	Average percent daytime Activity for subject i	$\frac{1}{m}\sum_{j=1}^{m}(\frac{\sum_{t=361}^{1080} log(AC_{ijt}+1)}{\sum_{t=1}^{1440} log(AC_{ijt}+1)} * 100)$
$\overline{Sed.Mins_i}$	Average sedentary minutes for subject i	$\frac{1}{m}\sum_{j=1}^{m}\sum_{t=1}^{1440} I(AC_{ijt<100})$
$\overline{MVPA.Mins_i}$	Average Moderate to Vigorous Active (MVPA) minutes for subject i	$\frac{1}{m}\sum_{j=1}^{m}\sum_{t=1}^{1440} I(AC_{ijt>2020})$
$\overline{wear.time_i}$	Average daily wear time for subject i	$\sum_{t=1}^{1440}(1-z_{ijk})$
$\mu_A(\mu_S)$	Average Active(Sedentary) Bout Duration	$\frac{T_A}{n_A}(\frac{T_S}{n_S})$
$g_A(g_S)$	Normalized Active(Sedentary) Bout Duration	$\frac{\sum_{t_1=1}^{D_A}\sum_{t_2=1}^{D_A} n_A(t_1)n_A(t_2)\lvert t_1-t_2\rvert}{2{n_A}^2\mu_A}$

Table 2.1 continued from previous page

$\bar{h_A}(\bar{h_r})$	Average Hazard from Active(Sedentary) to Sedentary(Active)	Suppose for observed durations t1,...,t_n, it is assumed that there are k unique values which are denoted in increasing order by $t_{n_1}, t_{n_2}, ...t_{n_k}$ $\bar{h_A} = \frac{1}{k}\sum_{t\in D}\frac{n_A(t_{n_i})}{n_A - n_A^c(t_{n_i-1})}$ $\bar{h_S} = \frac{1}{k}\sum_{t\in D}\frac{n_S(t_{n_i})}{n_S - n_S^c(t_{n_i-1})}$ where $D = t_{n_1}, t_{n_2}, ...t_{n_m}$
$\lambda_A(\lambda_S)$	Transition Probability from Active(Sedentary) to Sedentary(Active)	$\frac{n_A}{T_A}(\frac{n_S}{T_S})$
$\alpha_A(\alpha_S)$	Power Law Distribution	$1 + n_A[\sum_{i=1}^{n_A} ln\frac{t_i}{\hat{t_{min}}-0.5}]^{-1}$ $1 + n_s[\sum_{i=1}^{n_s} ln\frac{t_i}{\hat{t_{min}}-0.5}]^{-1}$

Table 2.2: Table 2.2 continued from previous page

Metrics	Notes
$\overline{TLAC_i}$	
$\overline{Pdaytime_i}$	Daytime is defined as 6am-6pm
$\overline{Sed.Mins_i}$	cut-point for sedentary activity is 100
$\overline{MVPA.Mins_i}$	cut-point for MVPA activity is 2020
$\overline{wear.time_i}$	
$\mu_A(\mu_S)$	
$g_A(g_S)$	Bounded between 0 and 1. When g is close to 1, it means the total active(sedentary) time is accumulated via a small number of longer bouts. When g is close to 0, it means all bouts contribute equally to total active(sedentary) time.
$\bar{h_A}(\bar{h_r})$	Larger values indicate a higher frequency of transitioning from active(sedentary) to sedentary (active) state.
$\lambda_A(\lambda_S)$	Larger values indicates more frequent transitioning between active (sedentary) states and sedentary (active) states.
$\alpha_A(\alpha_S)$	$\alpha_A(\alpha_S)$is the scaling parameter of the power law distribution, which is a distribution commonly used to model active(sedentary) bout duration. Larger values of,means that the total active(sedentary) time is accumulated via a larger number of shorter bouts.

Other demographic, co-morbidity, and lifestyle covariates considered in the analysis include: Age, Gender (binary), BMI Group (underweight, normal, overweight, obese), Race (White, Black, Others), Education (less than high school, high School, more than high school, missing education), Diabetes (binary), CHF (binary), CHD (binary), Cancer (binary), Stroke (binary), Smoking Status (current smoker, never smoker, former smoker), Alcohol (current drinker, never drinker, former drinker,missing alcohol status), Drinking Status (heavy drinker,moderate drinker, non-drinker, missing drinking status), Mobility Problem (binary), pulse in 60s (continuous), PulseIrregular (binary), difference in systolic blood pressure and diastolic blood pressure (continuous), Framingham Score (continuous) derived using information including age, gender, systolic blood pressure, current smoker or not, diabetes or not, total cholesterol, HDL-cholesterol, and blood pressure treated with medicine or not [4]. Of the participants included in our analysis, 2652 were alive after 5 years and 319 were deceased within 5 years. Table 2 compares the survey weights adjusted or unadjusted population characteristics between individuals who were alive and who were deceased within 5 years. Table 2 indicates that population characteristics adjusted for survey weights differ significantly between individuals who were alive and individuals who were deceased. We used a size of the test equal to $\alpha = 0.05$ level. The only variables that were not statistically different were race (p-value=0.1), and heart rate in 60s (p-value=0.107). Percent daytime activity is borderline significantly higher among alive individuals

(p-value=0.049). While the summary statistics from the survey weight adjusted analysis and unadjusted analysis are pretty consistent with each other, there are discrepancies for specific characteristics such as race, and percent daytime activity. The unadjusted analysis suggests that there are significantly higher proportion of white and a lower proportion of black individuals among the deceased individuals (p-value=0.001), and there is not enough evidence that the percent daytime activity differs between the alive and deceased individuals (p-value=0.136). These discrepancies may indicate the importance in incorporating survey weights into our analysis in order to obtain unbiased results.

Table 2.3: Comparison between survey weight adjusted/unadjusted population characteristics between individuals who were alive and individuals who were deceased within 5 years. Values in the table represent n(%) for categorical variables and mean(sd) for continuous variables

	Alive	Dead	p	Alive	Dead	p
n	2652	319		2695.2	275.8	
Gender = 1	1291 (48.7)	207 (64.9)	<0.001	1214.4 (45.1)	152.5 (55.3)	0.013
BMI category			<0.001			0.026
<18.5	23 (0.9)	9 (2.8)		28.7 (1.1)	8.6 (3.1)	
18.5-24.99	656 (24.7)	106 (33.2)		718.2 (26.6)	92.3 (33.5)	
25-29.99	1011 (38.1)	109 (34.2)		1017.5 (37.8)	91.6 (33.2)	

Table 2.3 continued from previous page

>30	962 (36.3)	95 (29.8)		930.8 (34.5)	83.3 (30.2)	
Race			0.001			0.1
White	1527 (57.6)	216 (67.7)		2153.1 (79.9)	231.1 (83.8)	
Black	497 (18.7)	54 (16.9)		257.5 (9.6)	25.7 (9.3)	
Other	628 (23.7)	49 (15.4)		284.6 (10.6)	19.1 (6.9)	
Education			<0.001			<0.001
LessThanHS	800 (30.2)	135 (42.3)		506.1 (18.8)	95.8 (34.7)	
High school	656 (24.7)	85 (26.6)		719.8 (26.7)	82.4 (29.9)	
MoreThanHS	1196 (45.1)	99 (31.0)		1469.3 (54.5)	97.7 (35.4)	
Diabetes = 1	431 (16.3)	84 (26.3)	<0.001	339.8 (12.6)	66.1 (24.0)	<0.001
CHF = 1	107 (4.0)	55 (17.2)	<0.001	97.5 (3.6)	53.4 (19.4)	<0.001
CHD = 1	184 (6.9)	50 (15.7)	<0.001	181.7 (6.7)	44.9 (16.3)	<0.001
Cancer = 1	373 (14.1)	79 (24.8)	<0.001	425.9 (15.8)	81.8 (29.7)	<0.001

Table 2.3 continued from previous page

Stroke = 1	127 (4.8)	47 (14.7)	<0.001	105.7 (3.9)	43.5 (15.8)	<0.001
SmokeCigs			<0.001			<0.001
Current	445 (16.8)	79 (24.8)		440.4 (16.3)	62.3 (22.6)	
Former	967 (36.5)	145 (45.5)		966.5 (35.9)	130.3 (47.3)	
Never	1240 (46.8)	95 (29.8)		1288.3 (47.8)	83.2 (30.2)	
Alcohol			<0.001			<0.001
Current	1490 (56.2)	133 (41.7)		1638.9 (60.8)	112.3 (40.7)	
Former	700 (26.4)	125 (39.2)		609.4 (22.6)	108.3 (39.3)	
MissingAlcohol	78 (2.9)	13 (4.1)		78.0 (2.9)	12.0 (4.3)	
Never	384 (14.5)	48 (15.0)		368.9 (13.7)	43.2 (15.7)	
DrinkStatus			<0.001			<0.001
HeavyDrinker	151 (5.7)	28 (8.8)		180.2 (6.7)	25.2 (9.1)	
MissingAlcohol	79 (3.0)	13 (4.1)		78.3 (2.9)	12.0 (4.3)	

Table 2.3 continued from previous page

ModerateDrinker	1338 (50.5)	105 (32.9)		1458.3 (54.1)	87.1 (31.6)	
Non-Drinker	1084 (40.9)	173 (54.2)		978.3 (36.3)	151.5 (54.9)	
MobilityProblem = 1	463 (17.5)	97 (30.4)	<0.001	424.7 (15.8)	88.0 (31.9)	<0.001
Pulse_irregular = 1	187 (7.1)	46 (14.4)	<0.001	186.4 (6.9)	37.6 (13.6)	0.001
Age	64.66 (9.14)	73.21 (8.96)	<0.001	63.12 (9.70)	73.71 (9.73)	<0.001
Pdaytime	75.49 (10.34)	74.57 (11.14)	0.136	75.19 (10.09)	73.83 (11.28)	0.049
TLAC	2811.59 (713.48)	2235.97 (757.64)	<0.001	2842.70 (693.60)	2214.02 (778.92)	<0.001
Sed_Mins_norm	60.94 (11.86)	71.05 (13.10)	<0.001	60.68 (11.50)	71.72 (12.76)	<0.001
MVPA_Mins_norm	1.68 (1.99)	0.71 (1.30)	<0.001	1.80 (2.03)	0.69 (1.38)	<0.001
μ_S	6.56 (2.75)	9.07 (6.24)	<0.001	6.48 (2.63)	9.38 (6.82)	<0.001
μ_A	3.89 (1.31)	3.07 (1.12)	<0.001	3.91 (1.29)	3.03 (1.06)	<0.001
λ_S	0.17 (0.05)	0.14 (0.06)	<0.001	0.17 (0.05)	0.13 (0.06)	<0.001
λ_A	0.28 (0.08)	0.36 (0.11)	<0.001	0.28 (0.08)	0.36 (0.11)	<0.001

Table 2.3 continued from previous page

$\bar{h_S}$	0.17 (0.04)	0.15 (0.04)	<0.001	0.17 (0.04)	0.15 (0.04)	<0.001
$\bar{h_A}$	0.27 (0.08)	0.36 (0.12)	<0.001	0.27 (0.08)	0.36 (0.12)	<0.001
g_S	0.62 (0.05)	0.64 (0.05)	<0.001	0.62 (0.05)	0.65 (0.05)	<0.001
g_A	0.50 (0.07)	0.45 (0.08)	<0.001	0.50 (0.06)	0.44 (0.08)	<0.001
α_A	1.65 (0.08)	1.72 (0.11)	<0.001	1.65 (0.08)	1.73 (0.10)	<0.001
α_S	1.57 (0.07)	1.52 (0.08)	<0.001	1.57 (0.07)	1.52 (0.08)	<0.001
systolic_ave	132.86 (20.65)	138.02 (26.07)	<0.001	131.24 (20.07)	137.78 (26.74)	0.001
diastolic_ave	70.41 (14.02)	65.18 (16.01)	<0.001	70.93 (13.82)	64.52 (16.38)	<0.001
Blood pressure diff	62.44 (22.13)	72.84 (25.69)	<0.001	60.31 (21.80)	73.26 (26.86)	<0.001
60sec pulse	70.01 (12.09)	71.45 (13.61)	0.047	69.99 (11.92)	71.21 (13.26)	0.107
Framingham score	14.35 (15.93)	22.11 (23.09)	<0.001	12.64 (14.22)	21.47 (21.99)	<0.001

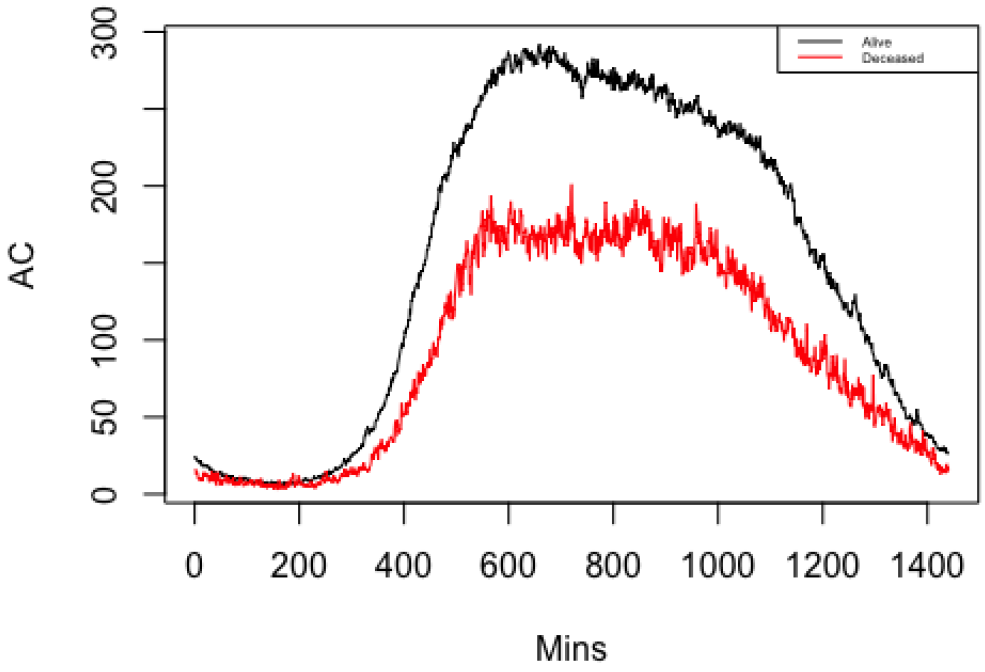
Chapter 3

Exploratory Data Analysis

To begin, we investigate the association between activity levels and mortality, exploratory analysis was conducted on the minute-level activity count data. We first computed the average minute-level activity count across days within subject. We then subsetted the data based on mortality status, and computed the mean activity count across subjects among those who were alive and those who were deceased after 5 years. Figure 3.1 displays the mean activity count trajectory in the two mortality groups. The black line denotes mean activity count trajectory among alive individuals, and the red line denotes mean activity count trajectory among deceased individuals. The figure indicates a visible difference between the activity trajectory of alive and deceased individuals. The alive individuals tend to have much higher average activity level, especially between 6am and 8pm.

Figure 3.2 displays the density distributions of the continuous variables, includ-

Figure 3.1: Mean activity count trajectory in the two mortality groups. Black line denotes mean activity count trajectory among alive individuals, and red line denotes mean activity count trajectory among deceased individuals



ing Age, Percent daytime activity (Pdaytime), Total Log Activity Count (TLAC),Sedentary Minutes normalized by the total wear time (Sed_Mins_norm), Moderate-to-Vigorous-Activity time normalized by the total wear time (MVPA_Mins_norm), the 10 fragmentation metrics, difference in systolic and diastolic blood pressure (BP diff), Pulse (heart rate) 60s,and Framingham Score, based on mortality status. From the plot, we noticed apparent shift in the density distribution of all the continuous variables, except for Pdaytime and Pulse 60s. For subjects who were alive, they tend to be younger, with higher TLAC, with less time spent in sedentary activities, more time spent in Moderate-to-Vigorous activities, shorter average sedentary bouts duration (μ_S), longer average active bouts duration (μ_A), higher average hazard of changing from being active to being sedentary ($\bar{h_A}$), higher transition probability of changing from being active to being sedentary (λ_A), lower average hazard of changing from being sedentary to being active ($\bar{h_S}$, lower transition probability from changing from being sedentary to being active (λ_S), less fragmented active time (g_A, α_A), more fragmented sedentary time (g_S, α_S), with a smaller difference between systolic and diastolic blood pressure, and a lower Framingham Score. The two-sided Wilcoxon Rank Sum Test further confirmed that the differences on the variables mentioned above are significant(p-value<0.001).

Figure 3.3 displays the correlation between the continuous variables. The correlations between Pdaytime and other activity related covariates are weak. There was a strong negative correlation between TLAC and Sed_Mins_norm, μ_S, λ_A, $\bar{h_A}$,

and α_A; Sed_Mins_norm and μ_A, λ_S,$\bar{h_S}$,g_A, and α_S ; μ and λ when the subscripts are the same; μ and α when the subscripts are the same; μ and $\bar{h}$ when the subscripts are the same; λ and g when the subscripts are the same; $\bar{h_A}$ and g_A; g_A and α_A. The directions of correlations between fragmentation metrics are consistent with what those reported in Junrui et. al [5]. This plot suggests that the high correlations between activity need to be taken into account during the variable selection process, which will be discussed later in section 4.

Figure 3.4 explores the potential interactions between TLAC and age as predictors of 5-year mortality. The colored lines display the smoothed observed probability of mortality versus age, stratified by different TLAC groups. TLAC was grouped based on quantiles, and they were clustered into 3 groups to denote <1st quantile, 1st to 3rd quantile, and >=3rd quantile. The smoothing is done using the generalized additive model (GAM) implemented in the 'mgcv' package[11]. The plot indicates that there may be a different relationship between mortality and age as a function of TLAC. Before age 65 being above the third quartile of TLAC or being between the first and third quartiles of TLAC does not make a large difference in the estimated trend of association between the probability of mortality by age. However, after age 65 the mortality probability for individuals with a TLAC between the first and third quartile is higher than for individuals with TLAC above the third quartile. The difference between the probability of mortality of individuals with TLAC above the third quartile and those with TLAC below the first quartile also increases substantially after age 65. This indicates that there may be interaction effects between TLAC and age when predicting 5-year mortality.

Figure 3.2: Density distributions of continuous variables based on mortality status

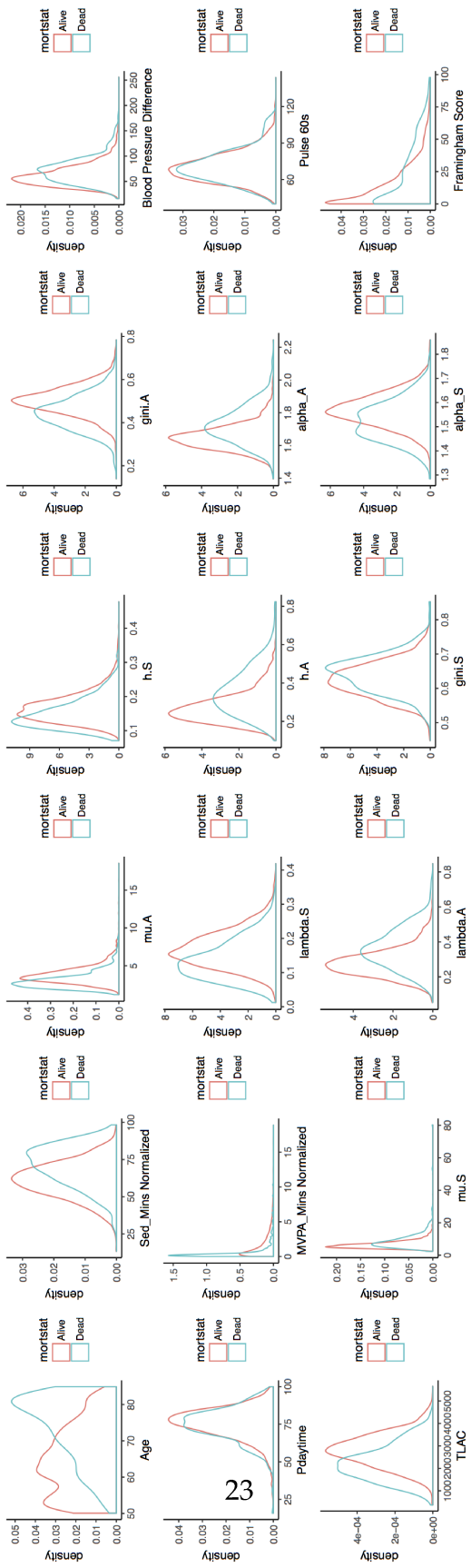


Figure 3.3: Explore the correlation between continuous variables

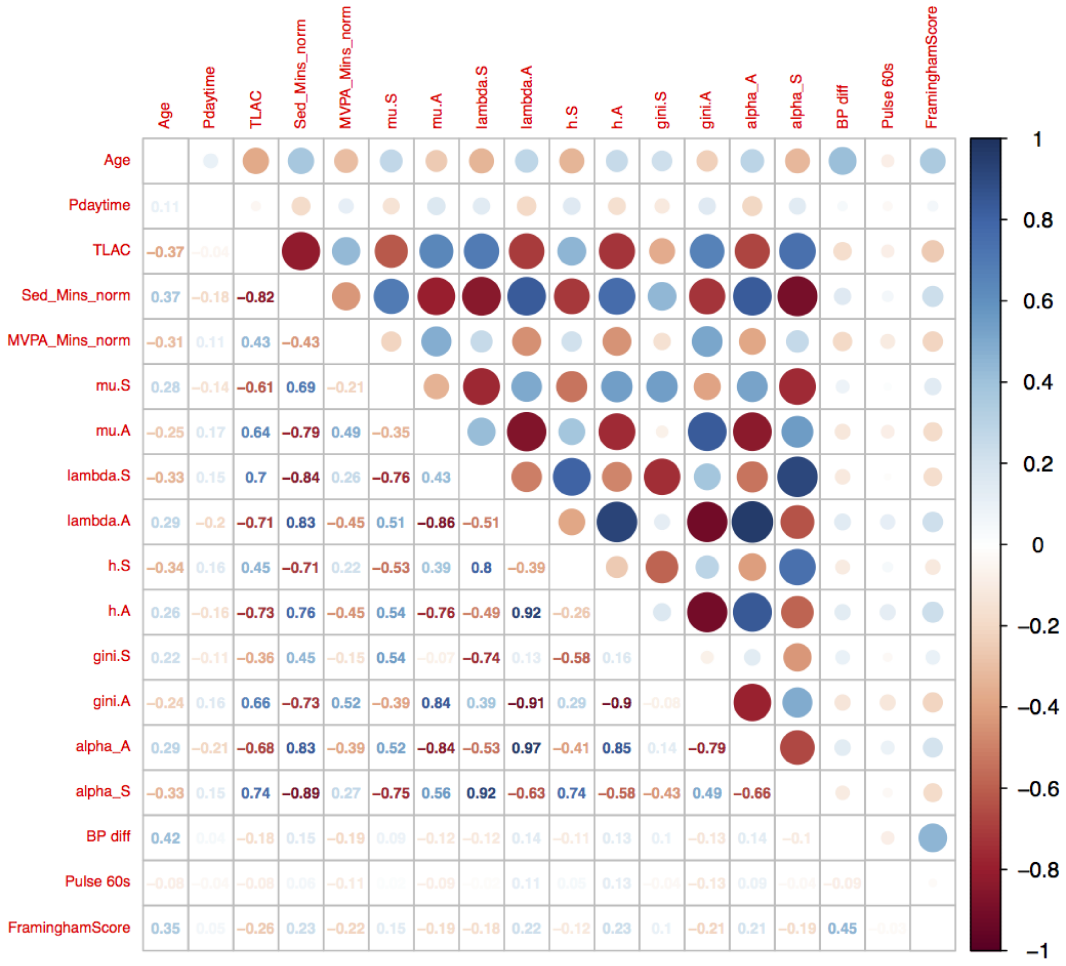
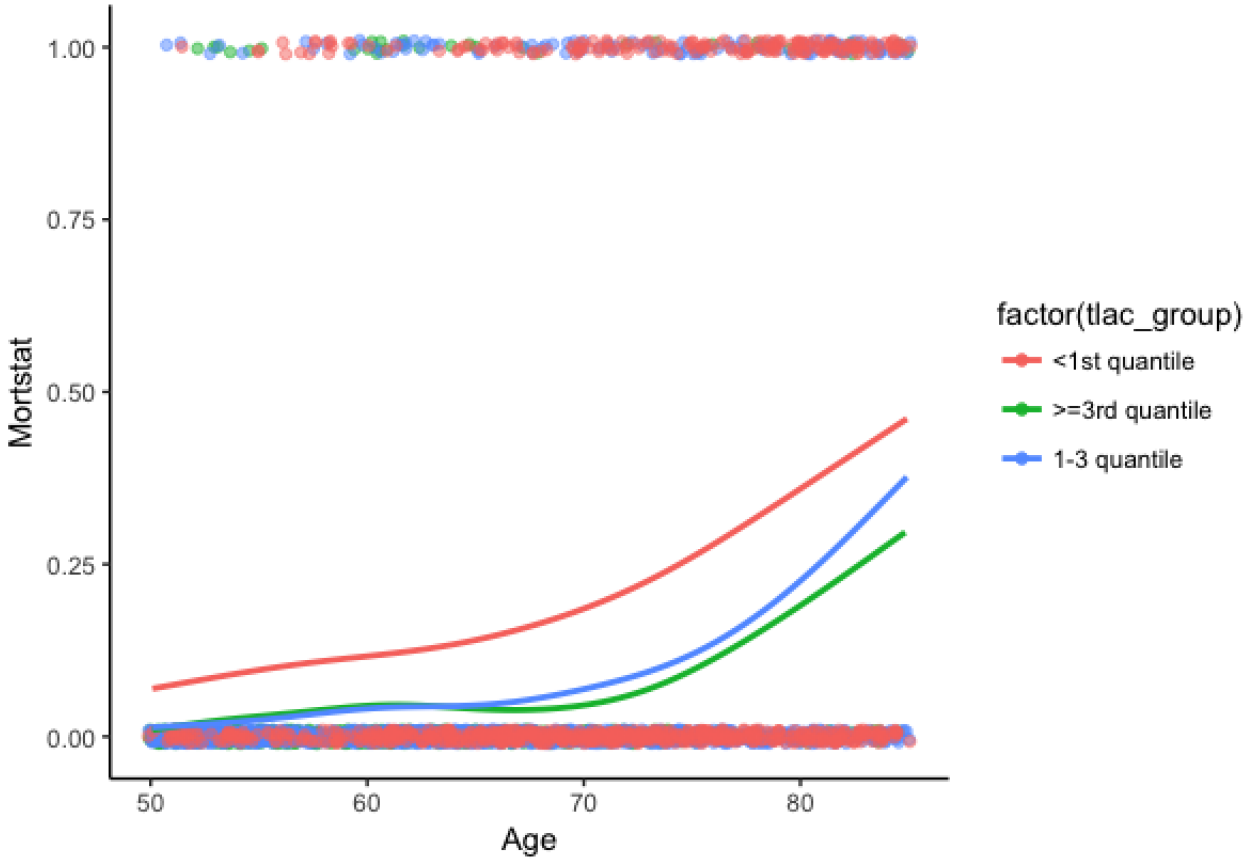


Figure 3.4: Explore interaction between TLAC and Age. The colored lines display the smoothed observed probability of mortality against age, stratified by different TLAC group



Chapter 4

Predictions

4.1 Model Selection

In this chapter we are interested in predictors of mortality in terms of their absolute and relative performance. First, we considered univariate logistic regression models with 5-year mortality as the outcome, and the activity metrics in Table 2.1, age, gender, race, education, BMI, mobility problem, CHF, diabetes, stroke, CHD, drinking status, alcohol status, smoking status, difference in systolic and diastolic blood pressure (BP Diff), pulse in 60s, PulseIrregular, and the Framingham Score as the predictor candidates. Each variable is ranked based on the 10-fold cross-validated area under receiver operation characteristics curve (AUC). Individual predictors ranking along with their estimated coefficient, t-statistics, and P-values for the univariate logistic regression model are listed in table 4.1. Table 4.1 indi-

cates that Age is the predictor that maximizes the cross-validated AUC (0.745), followed by three accelerometry-derived activity metrics: MVPA_Mins_norm, λ_A, and μ_A (AUC=0.725, 0.724,0.724, respectively). The absolute t-statistics for λ_A, $\bar{h_A}$, g_A, Sed_Mins_norm, and α_A are comparable to the t-statistic for the Age effect.

Table 4.1: Ranking of individual mortality predictors based on 10 cross-validated AUC

Variable	Coef	t stat	P-value	AUC
Age	0.103	14.113	0.000	0.745
MVPA_Mins_norm	-0.578	-8.289	0.000	0.725
λ_A	8.896	13.961	0.000	0.724
μ_A	-0.838	-11.221	0.000	0.724
$\bar{h_A}$	8.262	13.964	0.000	0.722
g_A	-12.176	-13.233	0.000	0.722
Sed_Mins_norm	0.074	13.194	0.000	0.720
α_A	8.805	13.330	0.000	0.713
TLAC	-0.001	-12.613	0.000	0.712
λ_S	-13.209	-10.325	0.000	0.678
μ_S	0.172	9.989	0.000	0.676
α_S	-9.555	-10.707	0.000	0.671
$\bar{h_S}$	-12.978	-7.554	0.000	0.642

Table 4.1 continued from previous page

BP Diff	0.017	7.428	0.000	0.627
gs	7.483	6.549	0.000	0.609
Framingham Score	0.022	7.486	0.000	0.566
Smoking Status				
Former Smoker	-0.169	-1.117	0.264	0.565
Never Smoker	-0.840	-5.188	0.000	0.565
Drinking Status				
MissingAlcohol	-0.119	-0.329	0.742	0.559
Moderate Drinker	-0.860	-3.749	0.000	0.559
Non-Drinker	-0.150	-0.678	0.498	0.559
Alcohol Status				
Former Drinker	0.693	5.224	0.000	0.558
MissingAlcohol	0.624	1.995	0.046	0.558
Never Drinker	0.337	1.893	0.058	0.558
Gender	0.667	5.398	0.000	0.558
Education				
High School	-0.264	-1.783	0.075	0.551
MoreThanHS	-0.712	-5.088	0.000	0.551
CHF	1.600	8.988	0.000	0.535

Table 4.1 continued from previous page

MobilityProblem	0.726	5.495	0.000	0.534
Race				
Black	-0.264	-1.642	0.101	0.527
Other	-0.595	-3.602	0.000	0.527
BMI Category				
18.5-24.99	-0.884	-2.174	0.030	0.526
25-29.99	-1.289	-3.176	0.001	0.526
>30	-1.377	-3.378	0.001	0.526
Diabetes	0.611	4.440	0.000	0.521
Pulse 60s	0.009	1.986	0.047	0.520
Cancer	0.699	4.948	0.000	0.516
CHD	0.914	5.314	0.000	0.514
Pdaytime	-0.008	-1.491	0.136	0.510
Stroke	1.234	6.771	0.000	0.508
Pulse Irregular	0.798	4.521	0.000	0.500

After conducting univariate regressions, we conducted forward selection on three subsets of mortality predictors: non-activity related metrics (age, gender, race, education, BMI, mobility problem, CHF, diabetes, stroke, CHD, drinking status, alcohol status, smoking status, BP Diff, pulse in 60s, PulseIrregular, and

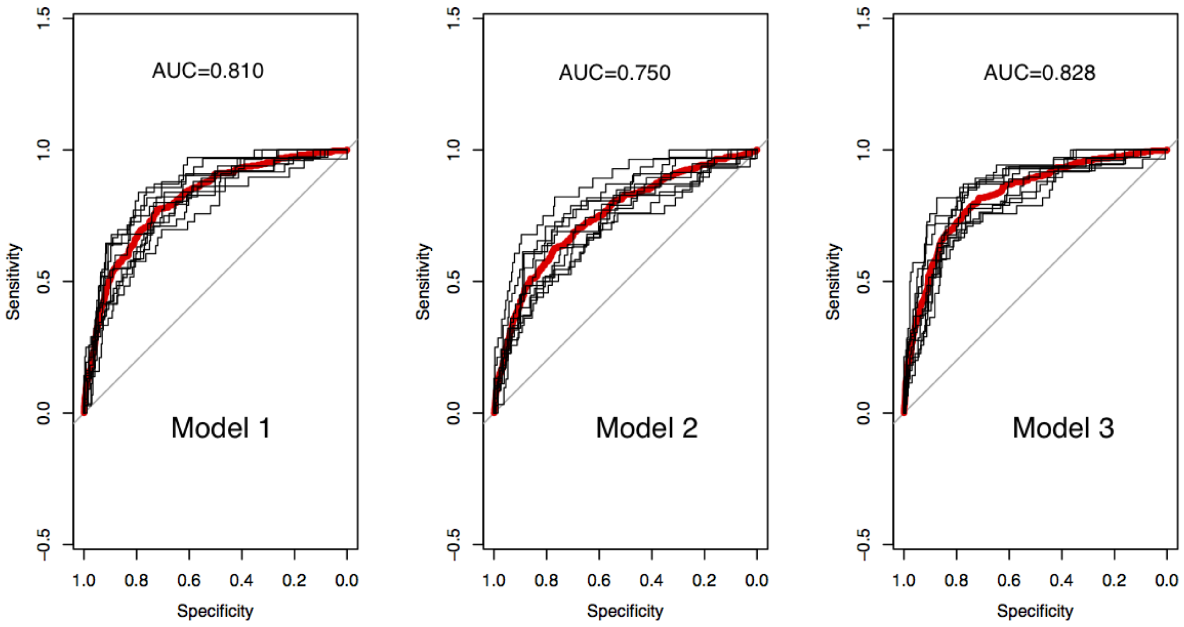
Framingham Score), activity metrics (as shown in table 2.1), and the combination of non-activity and activity metrics. We started by conducting forward selection based on non-activity related metrics and chose a benchmark model. The benchmark model was then used in a new forward selection procedure for activity metrics. Mortality predictors were added to the logistic regression model one at a time. At each step, each variable that is not already in the model is tested for inclusion in the model. The variable that maximizes the cross-validated AUC (or minimized the Akaike's Information Criteria (AIC) or the Effective Parsimony Information Criterion(EPIC)) is added to the model. To choose the benchmark model, we included all predictors that increased AUC (or decreased AIC/EPIC). EPIC and AIC are both Maximum Likelihood Estimate driven, however, EPIC penalizes free parameter more than AIC[8]. Therefore, AIC is less stringent than EPIC, and tends to accept more covariates into the model. Interestingly, the AUC-based criteria included more covariates in the benchmark model than both EPIC and AIC. When we conducted forward selection of non-activity related covariates, EPIC resulted in a model with 10 covariates, in the following order of Age: Smoking Status, CHF, Drinking Status, Gender, Pulse 60s, Diabetes, Stroke, Cancer, and Blood Pressure difference. AIC resulted in a model with 12 covariates, and also included Mobility Problem and BMI category in addition to those selected by EPIC. The AUC-based criterion resulted in a model with 13 covariates, and included CHD in addition to those selected by AIC. However, the AUC gain is quite small after including the

11th variable, and CHD is not significant when fitting the model (p-value=0.821). Therefore, we do not choose our benchmark model based on AUC.

To account for the high correlation between activity metrics, which we observed in Figure 3.3, we included an additional during the forward selection of activity metrics. Whenever an activity metric is selected as the predictor that maximizes the cross-validated AUC (or minimizes AIC/EPIC), the activity metric that are highly correlated with the selected activity metric (absolute correlation$>$ 0.6) are excluded from the candidate variable set. For example, if $\bar{h_A}$ is selected as the variable that maximizes the AUC (or minimizes AIC/EPIC), after $\bar{h_A}$ is added to the model, all the activity metrics that have absolute correlation with $\bar{h_A}$ greater than 0.6, including TLAC(corr=-0.73), Sed_Mins_norm(corr=0.76), μ_A (corr=-0.76), λ_A (corr=0.92), g_A (corr=-0.9), α_A (corr=0.85), are no longer considered as candidate predictors and are excluded from the forward selection process.

Based on the three subsets of mortality predictors, we chose three corresponding benchmark models based on EPIC, and the ROC curves for the three benchmark models were shown in Figure 4.1. The plots display each of the ROC curves for the 10 sub-samples from the cross-validation and gives their mean AUC. The mean ROC curves are shown in red. Model 1 was chosen based only on non-activity metrics and contains the following predictors: Age, Smoking Status, CHF, Drinking Status, Gender, pulse in 60s, Diabetes, Stroke, Cancer, and BP difference. The 10-fold cross validated AUC for this model was 0.81. Model 2 was cho-

Figure 4.1: ROC Curves for the Three Benchmark Model



sen based only on activity metrics and contained the following predictors: λ_A,g_S, MVPA_Mins_norm, and Pdaytime, and had a 10-fold cross-validated AUC as 0.750. Model 3 contained the predictors in Model 1 as well as $\bar{h_A}$ chosen based on forward selection, which increased the cross-validated AUC from 0.810 in Model 1 to 0.828.

4.2 Model Fitting

Model 3, which has the highest cross-validated AUC (0.828), was also fitted using survey weights and the estimated coefficients are shown in Table 4.2. We standardized the variable $\bar{h_A}$ by subtracting its mean and dividing by its standard

deviation for ease of interpretation. Table 4.2 indicates that the probability of 5 year mortality increases significantly with age, CHF, Cancer, being male, blood pressure difference, and average hazard from active to sedentary and decreases significantly if former smoker, never smoker, and moderate drinker. we estimate that the relative odds of mortality comparing an individual with CHF at baseline to an individual without CHF at baseline is exp(0.902)= 2.465 (95% CI:1.515, 4.011), the relative odds of mortality comparing males to females is exp(0.649)=1.914 (95% CI:1.314,2.786), and the relative odds of mortality comparing an individual with cancer at baseline to individual without cancer at baseline is exp(0.405)= 1.500 (95% CI:1.056, 2.130). For individuals with the same smoking status, CHF status, drinking status, gender, diabetes status, cancer status, pulse within 60s, blood pressure difference, and $\bar{h_A}$, a 1 year increase in age is associated with 9.9% increase in the odds of mortality; for individuals with the same age, smoking status, CHF status, drinking status, gender, diabetes status,cancer status, pulse within 60s, and $\bar{h_A}$, 1 unit increase in blood pressure difference is associated with 0.5% increase in the odds of mortality. FOr individuals of the same age, smoking status, CHF status, drinking status, gender, diabetes status, cancer status, pulse within 60s, and blood pressure difference, a 1 unit increase in standardized $\bar{h_A}$ is associated with 66.9% increase in odds of mortality. we estimate that the relative odds of mortality comparing former smoker with current smoker is exp(-0.745)= 0.474 (95% CI:0.279 0.807), the relative odds of mortality comparing never smoker with

current smoker is exp(-1.185)= 0.305 (95% CI:0.214,0.437), and the relative odds of mortality comparing moderate drinker with heavy drinker is exp(-0.922)= 0.398 (95% CI:0.203, 0.780). The Hosmer-Lemeshow goodness of fit test suggests lack of evidence against the null that the observed number of events is inconsistent with the predicted number of events (p-value=0.156).

In order to test for potential interactions between TLAC and age as predictors of 5-year mortality, we fitted a separate logistic regression model with 5-year mortality as the outcome, TLAC along with the non-activity covariates selected in Model 1 as the predictors. We employed likelihood ratio test to examine whether the interaction term between TLAC and Age significantly improve model fitting. However, after accounting for the other covariates the χ^2 test statistics failed to reject the null hypothesis that there is no interaction term between TLAC and Age (p-value=0.892).

Table 4.2: Estimated coefficients in the final,survey weighted, 8 predictors model

	Estimate	Std. Error	P-value	
Age	0.094	0.011	0.000	***
Smoking Status				
Former Smoker	-0.745	0.271	0.014	*
Never Smoker	-1.185	0.182	0.000	***
CHF	0.902	0.248	0.002	**
Drinking Status				
MissingAlcohol	-0.209	0.480	0.670	
Moderate Drinker	-0.922	0.344	0.016	*
Non-Drinker	-0.281	0.362	0.449	
Gender	0.649	0.192	0.004	**
pulse 60s	0.010	0.005	0.070	.
Diabetes	0.347	0.212	0.121	
Stroke	0.318	0.287	0.283	
Cancer	0.405	0.179	0.038	*
BP diff	0.005	0.002	0.044	*
$\bar{h_A}$	0.512	0.088	0.000	***

*Note: *p<0.05; **p<0.01; ***p<0.001*

Chapter 5

Survival Analysis and Competing Risk Model

We are also interested in assessing the effect of mortality predictors we selected in the previous forward selection process on heart disease and cancer related 5-year mortality, in addition to 5-year all cause mortality. In NHANES, individuals are also subject to other causes of death, including chronic lower respiratory diseases, Alzheimer disease, Diabetes and so on. Table 5.1 shows the 5-year mortality by cause of death. Among the individuals who are deceased within 5 years, 22.9% of deaths are due to heart disease, 28.3% are due to cancer, and 48.8% are due to causes other than heart disease and cancer. Here, death due to cancer and other causes are treated as competing events when heart disease related death is the outcome of interest. The idea is that as an individual who dies of a non-cardiovascular

Table 5.1: Summary of 5-year mortality by causes of death.

Cause of Death	n(%)
Heart Disease	73(22.9)
Cancer	90 (28.2)
Chronic lower respiratory diseases	21(6.6)
Cerebrovascular diseases	13(4.1)
Alzheimer's disease	7(2.2)
Diabetes	10(3.1)
Influenza and pneumonia	7(2.2)
Nephritis, nephrotic syndrome and nephrosis	8(2.5)
All other causes (residual)	90(28.2)
Total	319(100)

cause cannot die of heart disease. And similarly, death due to heart disease and other causes are treated as competing events when cancer related death is the outcome of interest.

Competing risk analysis is often used to analyze data that contains competing events. A common estimation approach is to estimate separately the probability of death for each type of event based on Kaplan-Meier(KM) product limit method. In this case, the other competing events are treated as censoring variables in addition to the times censored by withdrawal from the study or loss to follow-up.

For example, if the heart disease related death is the outcome of interest, we treat death due to all other causes as censoring, then apply the KM method to estimate the probability of death due to heart disease at each time point, as in standard survival analysis setting. This method of estimating event probabilities is called cause-specific hazard function,and can be expressed as:

$$\lambda_j(t) = lim_{dt \to 0+} \frac{P(t \leq T \leq t + dt, \pi = j | T \geq t)}{dt} ,$$

where T denotes time to death, π denotes cause of death, and $\pi = 1, 2, 3$ denotes death due to heart disease, death due to cancer, and death due to other causes in our setting.

Once this is done, we can investigate the association between the cause-specific hazard function and risk factors by modeling the cause-specific hazard function using, for example, the Cox proportional hazard model:

$$\lambda_j(t; z) = \lambda_{j0}(t) exp(z\beta_j)$$

However, as the cause-specific hazards model models hazard in the presence of competing risks, it may not provide causal interpretation for treatment or risk factors. An alternative method is based on cumulative incidence function, which is defined as

$$F_j(t) = P(T \leq t, \pi = j) = \int_0^t \lambda_j(u) S(u) du .$$

Here S(t) denotes the overall survival function at time t that can be estimated using

the standard KM estimato, and $\lambda_j(t)$ is the hazard for death due to cause j at time t, and can be estimated nonparametrically by $\frac{d_j(t)}{n(t)}$ where $d_j(t)$ denotes the number of deaths due to cause j at time t and $n(t)$ denotes the total number of observations at risk at time t. Then regression analysis can be done on the sub-distribution hazard function, which is defined as

$$\overline{\lambda_j(t)} = -\frac{d}{dt}log(1 - F_j(t)) = lim_{dt\to 0+}\frac{P(t \leq T \leq t + dt, \pi = j|T \geq t \cup (T \leq t \cap \pi \neq j))}{dt}.$$

The sub-distribution model can be expressed in the similar form as the cox regression model:

$$\overline{\lambda_j(t;z)} = \overline{\lambda_{j0}(t)}exp(z\beta_j)$$

[6]

However, the sub-distribution hazard is hard to interpret as it is unnatural. From the mathematical form of sub-distribution hazard, we can see that subjects who have already experienced a competing event remain in the risk set, although they are no longer at risk for the outcome of interest.

5.1 Model Fitting

We fit both cause-specific hazard model and sub-distribution hazard models to investigate the relative effects of predictors on mortality. We use the predictors that were selected for 5-year all-cause mortality and use them for heart disease and

cancer related 5-year mortality. Figure 5.2 displays the estimated hazard ratio with their 95% confidence intervals from the Cox proportional hazard model for 5-year all-cause mortality (All-cause mortality), cause-specific hazard model for heart disease related death (Heart Disease CSHM), and cause-specific hazard model when cancer related death is the outcome of interest (Cancer CSHM). We also display the sub-distribution hazard model for heart disease related death (Heart Disease SHM) and the sub-distribution hazard model for cancer related death (Cancer SHM). The Cox proportional hazard model for 5-year all cause mortality (red line) indicates that age, being diagnosed with CHF or cancer at baseline, being male, pulse 60s, blood pressure difference, and $\bar{h_A}$ significantly increase the all-cause mortality hazard. Being a former smoker, never smoker, and moderate drinker are significantly associated with decreasing the hazard for 5-year all cause mortality. The results are, in general, consistent with those from the survey-weighted logistic regression model, which we discussed in Chapter 4. When heart disease related death is the outcome of interest, moderate drinking, pulse 60s and cancer are no longer statistically significant. When cancer related death is the outcome of interest, CHF and blood pressure difference no longer reach statistical significance. The change in the statistical significance may simply be due to a loss of power given the smaller number of events and should not be necessarily interpreted as the evidence that these predictors are no longer important.

In addition, quitting smoking has much stronger effect in terms of risk reduc-

tion for heart disease related death compared with the risk for all-cause mortality. If all the other risk factors stay the same, the hazard for a former smoker to die due to heart disease is 0.254 relative to a current smoker (95% CI 0.138,9.468) according to Heart CSHM, and 0.272 (95%CI:0.15, 0.493) according to Heart SHM, while the hazard for a former smoker to die from any cause is 0.496 (95% CI:0.37, 0.666) relative to a current smoker. Therefore, quitting smoking reduces the risk for mortality, though it has a stronger effect on reducing the mortality risk from heart disease. Interestingly, the direction of the effect of moderate drinking and non-drinking flips when heart disease related death is the outcome of interest. However, this should not be over-interpreted as results are not statistically significant. This is largely due to the increased standard error of the estimate. Being male is associated with a much higher risk for heart disease related death, as the hazard for a male to die from heart disease is 2.52 (95%CI: 1.486, 4.272) according to Heart CSHM or 2.378 (95%CI:1.422, 3.975) according to Heart SHM, relative to the hazard for a female if all the other risk factors stay the same. However, the hazard for a male to die from any cause is 1.798 relative to that for a female (95%CI: 1.405, 2.301). Being diagnosed with CHF at baseline has much stronger effect in increasing the risk for heart disease related death [hazard ratio estimated from Heart CSHM: 2.843 (95% CI:1.536, 5.262) and hazard ratio estimated from Heart SHM: 2.411 (95%CI:1.247, 4.661)]. Being diagnosed with cancer at baseline has a much stronger effect on increasing the risk for cancer related death [hazard ratio estimated from Cancer

CSHM: 2.897 (95% CI:1.859, 4.516) and hazard ratio estimated from Cancer SHM: 2.887 (95%CI:1.82, 4.579)]. This makes sense as individuals who are diagnosed at baseline with a particular disease are more likely to die due to that particular disease. $\bar{h_A}$ has comparable effect on increasing risk for heart-disease related death, cancer related death, and all-cause mortality, although its effect is stronger when all-cause mortality is of concern.

Overall, the cause-specific hazard model and sub-distribution hazard model yield consistent results in terms of the direction and significance of the regression coefficients. However, the estimated effects of the regression coefficients tend to be larger when using the cause-specific hazard model. In addition, we noticed that the standard error for the estimated hazard ratio from Cancer CSHM, Cancer SHM, Heart Disease CSHM, and Heart Disease SHM are in general larger than those from the Cox Proportional Hazard Model with all-cause mortality as the outcome of interest. This is largely due to the reduction in number of primary events when we restrict to only one cause-specific death.

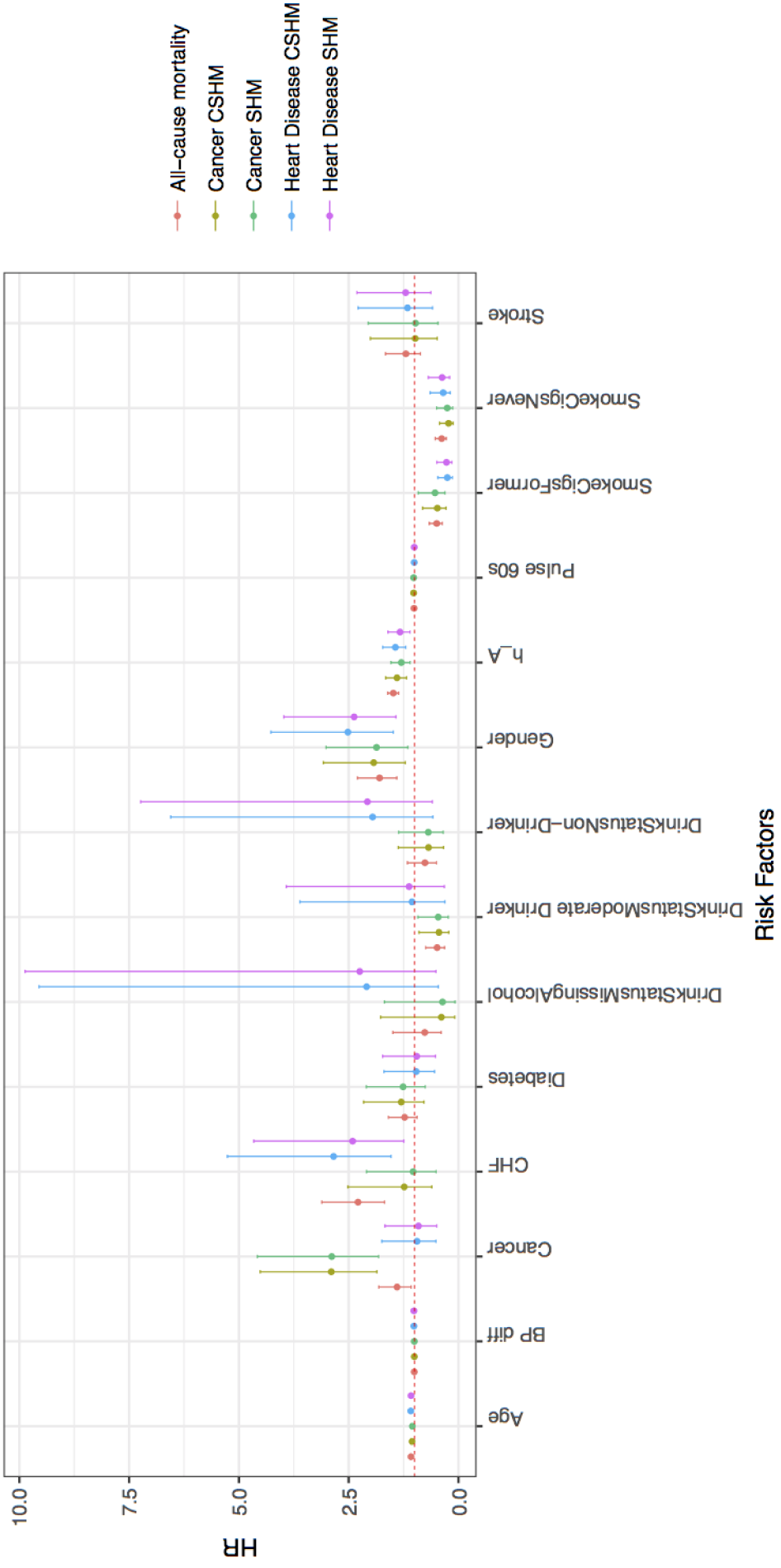


Figure 5.1: Comparison of estimated hazard ratio with their 95% confidence intervals from the Cox proportional hazard model for 5-year all cause mortality (All-cause mortality), cause-specific hazard model for heart disease related death (Heart Disease CSHM), cause-specific hazard model for cancer related death (Cancer CSHM), sub-distribution hazard model for heart disease related death (Heart Disease SHM), and sub-distribution hazard for cancer related death (Cancer SHM). The red dotted line denotes corresponds to a hazard rate HR=1.

5.2 Simulation

Austin et. al [1] concluded that an adequate number of primary events is required to accurately estimate the regression coefficients in the sub-distribution hazard model. In general, 40 to 50 events per variable (EPV) were necessary to ensure accuracy. However, when heart disease related death or cancer related death is the outcome of interest, the EPVs for some categorical variables fail to meet the $40 - 50$ criteria. Table 5.3 summarize the EPVs for each variable when heart disease or cancer related death is the outcome of interest. Table 5.3 indicates that for heart disease related death there are only 3 events among heavy drinkers, 22 events among moderate drinkers, and 16 events among cancer patients. For cancer related death there are only 9 events among CHF patients. In the sub-distribution hazard model for heart disease, moderate drinking and cancer no longer reach statistical significance though they show significant effects on all-cause mortality. In the sub-distribution hazard model for cancer, CHF does not reach statistical significance though it is significant in the all-cause mortality model. This is likely due to the smaller number of events, which is associated with a reduction in events per variable (EPV).

We proposed to use the upstrap [3] to examine the effect of EPV on the power of regression coefficients for the subdistribution hazard model. The detailed steps are listed as below:

Table 5.2: EPV for each variable in Heart Disease SHM and Cancer SHM

	Events Per Variable(EPV)	
	Heart Disease SHM	Cancer SHM
Age	73	90
Smoking Status		
Current Smoker	24	26
Former Smoker	23	45
Never Smoker	26	19
CHF	16	9
Drinking Status		
Heavy Drinker	3	11
MissingAlcohol	4	2
Moderate Drinker	22	32
Non-Drinker	44	45
Gender	49	61
Pulse 60s	73	90
Diabetes	17	23
Stroke	12	9
Cancer	13	34
BP diff	73	90
$\bar{h_A}$	73	90

1. Resample the dataset with replacement to 1, 1.5,..., 11 times more of the original dataset size

2. For each new resampled data, record the EPV for each variable, then fit the sub-distribution model for heart disease and cancer, record the p-values for each covariate, and repeat the process 1000 times for each size of the data

3. Compute the proportion of times p-value$\leq$ 0.05 as well as the mean EPV for each covariate across the 1000 upstrap samples with each given sample size

The proportion of times p-value$\leq$ 0.05 is the power to detect the effect of a covariate on the outcome of inteerst. Figure 5.2 and 5.3 display the power curve for each covariate in the sub-distribution hazard model for heart disease and cancer. The x-axis denotes the EPV for each covariate and the y-axis denotes the proportion of times the p-value$\leq$ 0.05. The horizontal red dotted line corresponds to power=0.6 and the vertical red dotted line corresponds to the EPV that corresponds to power=0.6.

As we expected, the power to detect the significance of each covariate increases as EPV increases. If we have around 150 events for non-drinking, which corresponds to a sample around 3.5 times larger than the original NHANES sample size, we will have approximately 60% power to detect the significance of non-drinking on the sub-distribution hazard for death from heart disease. If we have around 700 events for pulse 60s, which corresponds to a sample around 10 times larger

Figure 5.2: Power curves for each covariate in the sub-distribution hazard model for heart disease.

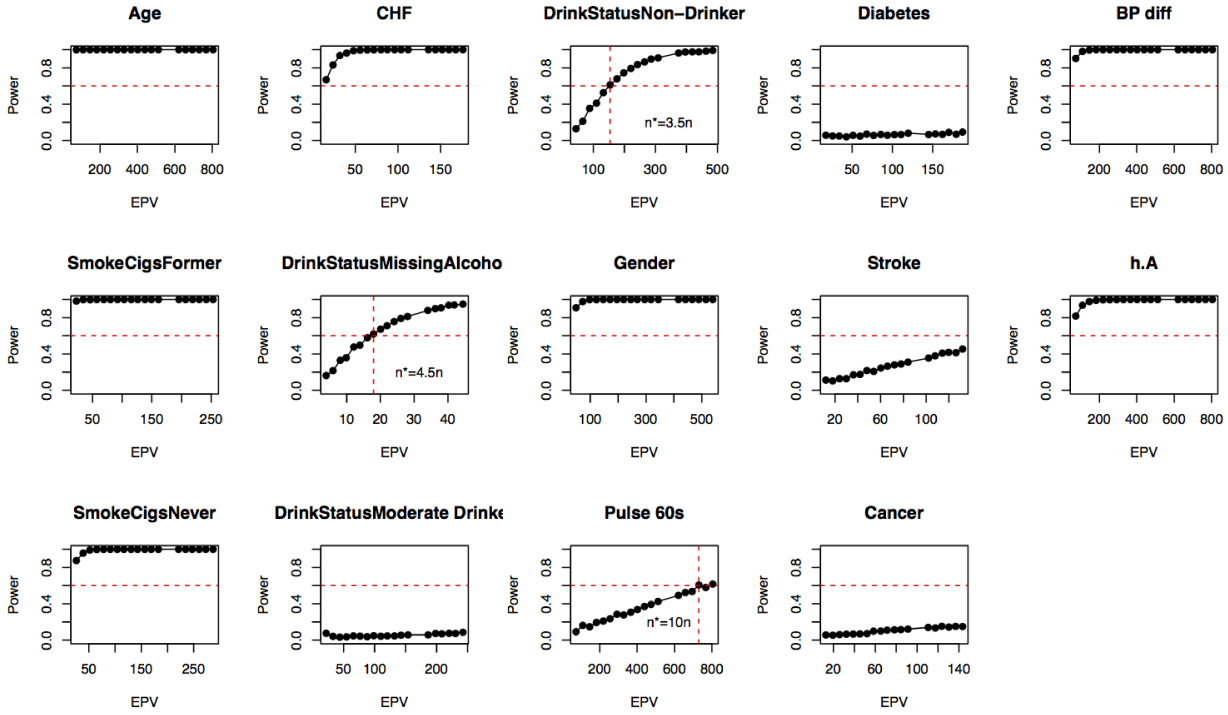
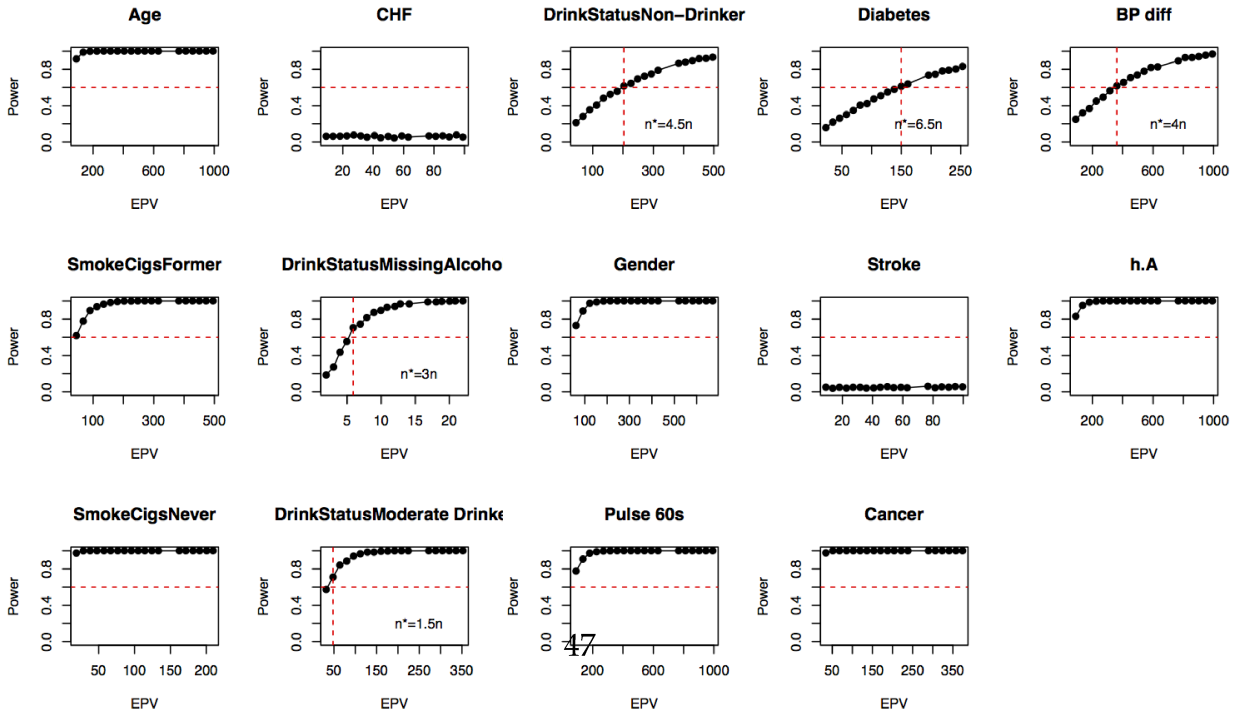


Figure 5.3: Power curves for each covariate in the sub-distribution hazard model for cancer.



than the original NHANES sample size, we will have approximately 60% power to detect the significance of Pulse 60s on the sub-distribution hazard for death from heart disease. Similarly, if we have around 400 events for blood pressure difference, which corresponds to a sample around 4 times larger than the original NHANES sample size, we will have approximately 60% power to detect the significance of blood pressure difference on the sub-distribution hazard for death from cancer. However, we still do not have enough power to detect the significance of some covariates, such as moderate drinking and baseline cancer status on the the sub-distribution hazard for death from heart disease, as well as baseline CHF status and baseline stroke status on the sub-distribution hazard for death from cancer, even after expanding the sample to 11 times of the original NHANES sample size. This provides a new perspective on the sample size that would be necessary to detect certain effects in relatively complex modeling scenarios.

Chapter 6

Mortality Prediction

In order to better visualize the probability of 5 year mortality for individuals with different risk factors and provide better individualized advice, we made a shiny APP that can be found at `https://yezheng42.shinyapps.io/MortalityCalculator/`.

This Shiny APP takes a users age, gender, pulse for 60s, comorbidity information including whether the person has CHF, cancer, stroke or diabetes at current age, blood pressure, self-reported activity percentile among peers, and smoking and drinking status. Using this information it estimates the predicted probability of all-cause mortality within 5 years, as well as the predicted risk relative to a reference person, who has the same non-modifiable risk factors but has median activity percentile of his age, is a never smoker and a non-drinker. The prediction is based on logistic regression that includes Age, gender, CHF, Cancer, Diabetes, Stroke, smoking status, drinking status, and TLAC as the predictors. The Hosmer-

Lemeshow goodness of fit test suggests lack of evidence against the null that the prediction model is not calibrated (p-value=0.602). If the user knows his or her blood pressure or pulse information, the prediction will take that information into account and will be built on the extended logistic regression model with further adjustments for pulse 60s and blood pressure difference. The Hosmer-Lemeshow goodness of fit test on the extended model also suggests lack of evidence against the null that the prediction model is not calibrated (p-value=0.165). If the comorbidity status is unknown by the user, then the model assumes that she or he does not have the comorbidity. The regression coefficients are estimated from our analysis dataset, which has 2971 observations. In these models we used TLAC instead of $\bar{h_A}$ as our activity metric, since it is easier to translate TLAC into activity percentile for different age groups. To make the translation, we used the empirical cumulative density function (CDF) of TLAC for the age group, and imputed the p^{th} percentile for the age group into the prediction equation. Age groups are defined in 10 year increments starting at 50 and these age groups are only used for activity; age enters the prediction model as a continuous predictor. For example, if a 55 year old individual reports her or his activity at the 25th percentile among their peers, then we apply the following procedure. First, we estimate the empirical CDF of TLAC for the age group between 50 and 60 years of age and estimate the 0.25 quantile of this distribution. This value is then plugged into the mortality risk prediction equation. The shiny app allows individuals to adjust their modi-

fiable health risk factors, including their activity percentile relative to their peers, smoking status, and drinking status. The software provides immediate feedback on how their trajectory of 5-year mortality risk is changed as a function of these factors and age.

Chapter 7

Discussion and Conclusion

In this thesis, we investigated the associations between participants' physical activity, demographic, and health-related characteristics with 5-year all-cause mortality using the data from the NHANES 2003-2004 and 2005-2006. We mainly focused on older individuals (age between 50-85 y.o.), and we identified that Age, Smoking Status, CHF, Drinking Status, Gender, Pulse 60s, Diabetes, Stroke, Cancer, Blood Pressure difference, and fragmentation metric $\bar{h_A}$ are the most predictive predictors for 5-year all-cause mortality. Together, these risk factors were used to build a prediction model, which has a 10-fold cross-validated AUC equal to 0.828. Using the Cox Proportional Hazard Model with 5-year all-indicated that: 1) age, being diagnosed with CHF or cancer at baseline, being male, having high heart rate, having high blood pressure difference, and having a higher transition probability from active to sedentary are significantly associated with increasing the haz-

ard for 5-year all-cause mortality; and 2) being a former smoker, never smoker, and moderate drinker are significantly associated with decreasing the hazard for 5-year all cause mortality. Our competing risk analyses for heart disease and cancer related deaths supported our findings for all cause mortality and provided further insights into cause-specific predictors. In particular, moderate drinking, heart rate, and having cancer were no longer identified as strong predictors of risk of death from heart disease. Moreover, CHF and having a larger blood pressure difference were no longer identified as strong predictors of risk of death from cancer. These findings are most likely due to the decrease in number of events from all cause to cause specific mortality data. Nevertheless, these findings are consistent with what we would expect in these scenarios. To get a better idea about the sample sizes at which various variables would become significantly associated with the hazard of death we conducted an upstrap re-sampling analysis. In general, the power to detect the significant effect for each covariate goes up as the sample size and EPV increases. Our proposed re-sampling method has the potential to estimate sample size, not only in the competing risk setting, but in any type of analysis regardless of their complexity. Moreover, we built a Shiny APP to translate our work to the general public. The APP predicts the probability of all-cause mortality within 5 years, as well as the predicted risk relative to a reference person after user inputs the required information. Users can also visualize how their predicted probability of all-cause mortality can change over time by modifying their modifiable risk

factors (such as quitting smoking or drinking less or exercising more).

There are some limitations of this study that should be kept in mind when interpreting the results. First, we are using cross-sectional data in NHANES, which cannot truly be used to infer the effects of changing modifiable risk factors within-person longitudinally. Information on variables were collected only at baseline and thus changes in variables (drinking status, blood pressure difference, smoking status, etc) cannot be analyzed. The prospective data in NHANES comes only from patients' mortality status, which are retrieved from the National Death Index. Second, some of the information on variables were collected based on self-reported data, and, thus, misclassification could be a serious problem, which may induce both bias and measurement error in the predictor variables. Third, as with any observational study, the problem of residual confounding from variables not measured in the study or inaccurate measurement or categorization of variables cannot be ruled out. However, since we have focused primarily on prediction, confounding is less important in this thesis.

The major merits of our study is that the NHANES sample is representative of the U.S. population and that the sample size is large enough to evaluate a wide range of mortality risk predictors. Our analysis adds to the body of evidence demonstrating the beneficial health effects of physical activity, non-smoking, and moderate drinking.

Bibliography

[1] Peter C. Austin, Arthur Allignol, and Jason P. Fine. "The number of primary events per variable affects estimation of the subdistribution hazard competing risks model". In: *Journal of Clinical Epidemiology* 83 (2017), pp. 75–84. ISSN: 18785921. DOI: 10.1016/j.jclinepi.2016.11.017.

[2] Centers for Disease Control and Prevention. *About the National Health and Nutrition Examination Survey*. 2016. URL: {http://www.cdc.gov/nchs/nhanes/about_nhanes.htm}.

[3] Ciprian Crainiceanu. "The upstrap". In: *bioRxiv* (2018). DOI: 10.1101/262436. URL: https://www.biorxiv.org/content/early/2018/02/15/262436.

[4] Ralph B. D'Agostino et al. "General cardiovascular risk profile for use in primary care: The Framingham heart study". In: *Circulation* 117.6 (2008), pp. 743–753. ISSN: 00097322. DOI: 10.1161/CIRCULATIONAHA.107.699579.

[5] Junrui Di et al. "Patterns of sedentary and active time accumulation are associated with mortality in US adults: The NHANES study". In: *bioRxiv* (2017).

DOI: https://doi.org/10.1101/182337. URL: http://biorxiv.org/content/early/2017/08/31/182337.abstract.

[6] Jason P. Fine and Robert J. Gray. "A Proportional Hazards Model for the Subdistribution of a Competing Risk". In: *Journal of the American Statistical Association* 94.446 (1999), pp. 496–509. ISSN: 1537274X. DOI: 10.1080/01621459.1999.10474144.

[7] Andrew Leroux. *nhanesdata: NHANES Accelerometry Data Pipeline*. R package version 1.0. URL: https://github.com/andrew-leroux/nhanesdata.

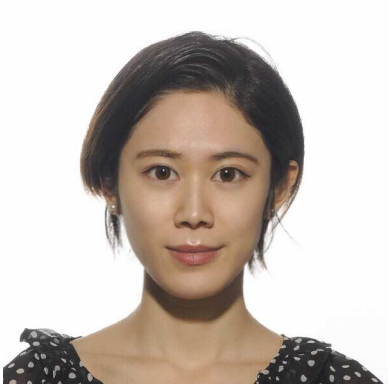
[8] Andrew Leroux et al. "Organizing and analyzing the activity data in NHANES".

[9] National Center for Health Statistics. *Office of Analysis and Epidemiology, Public-use Linked Mortality File*. Hyattsville, Maryland, 2015.

[10] R P Troiano et al. "Physical activity in the United States measured by accelerometer". In: *Medicine & Science in Sports & Exercise* 40.1 (2008), pp. 181–188.

[11] Simon Wood. *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation and GAMMs by REML/PQL*. R package version 1.8-20, 2017.

Vita



Yeya Zheng received the BS degree in Dietetics and Statistics from University of Wisconsin-Madison in 2016, and enrolled in the Biostatistics ScM program at Johns Hopkins University in 2016. She developed particular interest in wearable computing and joined the Wearable and Implantable Technology research group in 2016. During her first year of the master program, she worked with scientists of the Early Infant Care and Risk of Obesity study to learn how physical activity affects the growth of newly born babies. Starting in August 2018, Yeya will join Analysis Group, Inc as a healthcare analyst, where she will bring her statistical skills to help organizations improve the quality of care and enhance the patient experience.